AD-A037 559   COLORADO STATE UNIV FORT COLLINS DEPT OF MATHEMATICS   F/G 12/1
BEST RATIONAL APPROXIMATIONS WITH NEGATIVE POLES TO E TO THE -X--ETC(U)
DEC 76 E H KAUFMAN, G D TAYLOR   AF-AFOSR-2878-76
UNCLASSIFIED   AFOSR-TR-77-0166   NL

1 OF 1
AD A037559

END
DATE
FILMED
4-77

BEST RATIONAL APPROXIMATIONS WITH NEGATIVE POLES TO e^{-x} ON $[0,\infty)$

E.H. Kaufman, Jr. and G.D. Taylor

In this paper a theory for approximating e^{-x} on $[0,\infty)$ with rational functions having negative poles is developed. Numerical results suggest that the best uniform approximation to e^{-x} on $[0,\infty)$ from this class has only one pole and this is shown to be the case when using rational functions of this form which are linear polynomials divided by quadratic polynomials. Numerical results are given and compared to recent results of Saff, Schönhage and Varga.

1 Introduction

Let π_m denote the space of all real algebraic polynomials of degree less than or equal to m. For each m = 1,2,..., define R_m by

$$R_m = \{R=P/Q:\ P \in \pi_{m-1},\ Q(x)=\prod_{i=1}^{m} (q_i x+1),\ q_i \geq 0 \text{ for all } i\}.$$

Thus, R_m is the collection of all rational functions with negative poles from $R_m^{m-1}[0,\infty)$. Define λ_m by

$$(1.1)\quad \lambda_m = \inf\{||e^{-x} - R||_{L^\infty[0,\infty)}:\ R \in R_m\}.$$

It is known that λ_m converges geometrically to zero (i.e. $\overline{\lim_{m\to\infty}} \lambda_m^{1/m} = 0$) since Saff, Schönhage and Varga [5] have proved that there exists a sequence $\{R_m\}_{m=1}^{\infty}$, with $R_m(x) = P_{m-1}(x)/(1+\frac{x}{m})^m$, $P_{m-1} \in \pi_{m-1}$ such that

$$3 - 2\sqrt{2} \leq \overline{\lim_{m\to\infty}}\ ||e^{-x} - R_m||_{L^\infty[0,\infty)}^{1/m} \leq \frac{1}{2}.$$

In addition, since the poles of $R_m(x)$ are all real it follows that $R_m(z)$ must converge geometrically to e^{-z} in an infinite sector

BEST RATIONAL APPROXIMATIONS WITH NEGATIVE POLES TO e^{-x} ON $[0,\infty)$

E.H. Kaufman, Jr. and G.D. Taylor

In this paper a theory for approximating e^{-x} on $[0,\infty)$ with rational functions having negative poles is developed. Numerical results suggest that the best uniform approximation to e^{-x} on $[0,\infty)$ from this class has only one pole and this is shown to be the case when using rational functions of this form which are linear polynomials divided by quadratic polynomials. Numerical results are given and compared to recent results of Saff, Schönhage and Varga.

1 Introduction

Let π_m denote the space of all real algebraic polynomials of degree less than or equal to m. For each $m = 1,2,\ldots$, define R_m by

$$R_m = \{R=P/Q:\ P \in \pi_{m-1},\ Q(x) = \prod_{i=1}^{m}(q_i x+1),\ q_i \geq 0 \text{ for all } i\}.$$

Thus, R_m is the collection of all rational functions with negative poles from $R_m^{-1}[0,\infty)$. Define λ_m by

$$(1.1)\quad \lambda_m = \inf\{\|e^{-x} - R\|_{L^\infty[0,\infty)} : R \in R_m\}.$$

It is known that λ_m converges geometrically to zero (i.e. $\overline{\lim}\,\lambda_m^{1/m} < 1$) since Saff, Schönhage and Varga [5] have proved that there exists a sequence $\{R_m\}_{m=1}^{\infty}$, with $R_m(x) = P_{m-1}(x)/(1+\frac{x}{m})^m$, $P_{m-1} \in \pi_{m-1}$, such that

$$3 - 2\sqrt{2} \leq \overline{\lim_{m\to\infty}}\,\{\|e^{-x} - R_m\|_{L^\infty[0,\infty)}\}^{1/m}$$

In addition, since the poles of $R_m(x)$ are all at $-m$, $R_m(x)$ must converge geometrically to e^{-x} on an infinite

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER 19 18 AFOSR - TR - 77 - 0166	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) 6 BEST RATIONAL APPROXIMATIONS WITH NEGATIVE POLES to e^{-x} ON $[0, \infty)$		5. TYPE OF REPORT & PERIOD COVERED 9 Interim rept. 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) 10 E.H. Kaufman, Jr. [redacted] G.D. Taylor		8. CONTRACT OR GRANT NUMBER(s) 15 AF-AFOSR-[redacted]2878-76
9. PERFORMING ORGANIZATION NAME AND ADDRESS Colorado State University Department of Mathematics Fort Collins, Colorado 80523		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61102F 16 2304 17 A2
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Office of Scientific Research/NM Bldg. 410, Bolling AFB Washington, D. C. 20332		12. REPORT DATE 11 Dec[redacted]76 13. NUMBER OF PAGES 13 12 15p.
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release, distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

407 344

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Constrainted rational approximation to e^{-x}, numerical solution of parabolic partial differential equations via semi-discretization.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

In this paper a theory for approximating e^{-x} on $[0, \infty)$ infinity with rational functions having negative poles is developed. Numerical results suggest that the best uniform approximation to e^{-x} on $[0, \infty)$ from this class has only one pole and this is shown to be the case when using rational functions of this form which are linear polynomials divided by quadratic polynomials. Numerical results are given and compared to recent results of Saff, Schönhage and Varga.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered) symmetric about the positive x-axis [6]. In what follows, we define

$$(1.2)\quad \mu_m = \inf\{||e^{-x} - \frac{P(x)}{(1+\frac{x}{m})^m}||_{L^\infty[0,\infty)}: P \varepsilon \pi_{m-1}\}.$$

An application of this theory is in the construction of numerical algorithms for solving linear systems of ordinary differential equations which arise from semi-discretization of linear parabolic partial differential equations (see [1], [5]). Numerically, this reduces to an iteration of the form

$$(1.3)\quad \underset{\sim}{w}^{(r)} = A^{-1}\underset{\sim}{k} + R_m(\Delta tA)\{\underset{\sim}{w}^{(r-1)} - A^{-1}\underset{\sim}{k}\}$$

where A is a nxn matrix (band), $\underset{\sim}{k}$ and $\underset{\sim}{w}^{(j)}$ are n dimensional vectors (n is related to the stepsize of the discretization), Δt is a scalar and R_m is the rational function defined above. Due to the special form of the denominator of R_m, $\underset{\sim}{w}^{(r)}$ can be obtained from the repeated inversion of $(I + \frac{\Delta t}{m} A)\underset{\sim}{g}_{\ell+1} = \underset{\sim}{g}_\ell$, $0 \leq \ell \leq m - 1$ using an appropriately defined $\underset{\sim}{g}_0$. This is an attractive method numerically, since an LU factorization can be done for $I + \frac{\Delta t}{m} A$ only once and this factorization will preserve any band structure that is present.

One can construct a similar numerical method using a solution $R^*_m(x) = P^*_{m-1}(x)/\prod_{i=1}^{m}(q_ix+1)$ to (1.1). The apparent disadvantage of such a method compared to that of [5] is that $\underset{\sim}{w}^{(r)}$ is found from

$$\{\prod_{i=1}^{m}(I+q_i\Delta tA)\}\underset{\sim}{w}^{(r)}=\{\prod_{i=1}^{m}(I+q_i\Delta tA)\}A^{-1}\underset{\sim}{k}+P^*_{m-1}(\Delta tA)\{\underset{\sim}{w}^{(r-1)}-A^{-1}\underset{\sim}{k}\}$$

which will involve a greater number of operations (though, less than m LU factorization and 2m substitutions). The advantage of this method is that λ_m will be smaller than μ_m giving increased accuracy. However, it appears (numerically) that R^*_m actually has $q_1 = q_2 = \ldots = q_m$ and λ_m is approximately one half of μ_m for $m > 2$. Thus, using R^*_m gives a method that has the same desirable properties as that of (1.3) and increased accuracy.

In the next section we shall state some general facts concerning uniform approximation from R_m (these results will appear in a future paper [3]), give a theoretical treatment of best uniform approximation of (1.1) for the special case that $m = 2$ and state some conjectures. In the last section we will discuss our algorithm and present some numerical results.

2 Theoretical Results

In this section we begin by giving a existence theorem for approximating from R_m on $[0,\infty)$. This result is valid for a large class of functions (containing e^{-x}). We shall outline a proof for the special case that $m = 2$.

THEOREM 2.1. There exists $R^* \varepsilon R_m$ for which $||e^{-x}-R^*(x)||_{L^\infty[0,\infty)} = \lambda_m$.

THEOREM 2.2. There exists $R^* \varepsilon R_2$ for which $||e^{-x}-R^*(x)||_{L^\infty[0,\infty)} = \lambda_2$.

Proof. The proof begins by first observing that $\lambda_2 < \frac{1}{2}$. Thus, let $\{a_n\}$, $\{b_n\}$, $\{q_{1n}\}$ and $\{q_{2n}\}$ $(n=1,2,\ldots)$ be sequences such that $q_{1n} \geq 0$, $q_{2n} \geq 0$ for all n and $\frac{1}{2} \geq ||e^{-x} - \frac{a_n x+b_n}{(q_{1n}x+1)(q_{2n}x+1)}|| \downarrow \lambda_2$ as $n \to \infty$, where we will no longer write the subscript $L^\infty[0,\infty)$ on the norm bars. Next, the proof is divided into two cases. The first case is when the sequences $\{q_{1m}\}$ and $\{q_{2n}\}$ are bounded. In this case it follows that the sequences $\{a_n\}$ and $\{b_n\}$ are also bounded and the desired result follows as in the standard rational approximation theory.

Thus, let us assume that the sequences $\{q_{1n}\}$ and $\{q_{2n}\}$ are not both bounded. By relabelling and extracting subsequences we may assume that $q_{1n} \uparrow \infty$ and $q_{1n} \geq q_{2n}$ for all n. In addition, by looking at the error curve at $x = 0$ and $x = \frac{1}{2}$, respectively, we have that $\frac{1}{2} \leq b_n \leq \frac{3}{2}$ for all n, and for all n sufficiently large (say $n \geq n_o$) that

$$(2.1) \quad 0 < \eta \leq \frac{\frac{1}{2} a_n}{(\frac{1}{2} q_{1n}+1)(\frac{1}{2} q_{2n}+1)} \leq M$$

where η and M are positive constants independent of n.

Next, we claim that the sequence $\{q_{2n}\}$ must be bounded. Indeed, if not then by passing to subsequences (and relabelling) we may assume that $q_{2n} \to \infty$ as $n \to \infty$. Let $x_n = \left[\eta(\frac{1}{2}q_{1n}+1)(\frac{1}{2}q_{2n}+1)\right]^{-1}$ $\epsilon\ [0,\infty)$. Note that $x_n \downarrow 0$ and that

$$\frac{a_n x_n + b_n}{(q_{1n}x_n+1)(q_{2n}x_n+1)} \geq \frac{a_n x_n}{(q_{1n}x_n+1)(q_{2n}x_n+1)} \to 2$$

as $n \to \infty$. For n sufficiently large this contradicts our assumption that $||e^{-x} - \frac{a_n x + b_n}{(q_{1n}x+1)(q_{2n}x+1)}|| \leq \frac{1}{2}$. Thus, $\{q_{2n}\}$ must be bounded. Since $a_n \to \infty$ by (2.1), we have, reciprocating (2.1), that there exists positive constants c_1 and c_2 independent of n such that for $n \geq n_1 \geq n_o$,

$$(2.2) \quad \frac{c_1}{\frac{1}{2} q_{2n}+1} \leq \frac{q_{1n}}{a_n} \leq \frac{c_2}{\frac{1}{2} q_{2n}+1}.$$

By (2.2) we may extract a subsequence (and relabel) for which $q_{2n} \to q^* \geq 0$ and $q_{1n}/a_n \to c^* > 0$ as $n \to \infty$. Hence, for fixed $x\ \epsilon\ (0,\infty)$ we have that

$$\frac{a_n x + b_n}{(q_{1n}x+1)(q_{2n}x+1)} \to \frac{1}{c^*(q^*x+1)} =: \frac{b^*}{q^*x+1} .$$

By continuity, $||e^{-x} - \frac{b^*}{q^*x+1}|| \leq \lambda_2$ completing the argument.∎

The above proof (suitably modified) also establishes the following corollary where $\tilde{R}_m = \{R=P/Q\colon P \in \pi_{m-1}, Q(x)=(qx+1)^m, q \geq 0\}$.

COROLLARY 2.3. <u>There exists</u> $\hat{R} \in \tilde{R}_m$ <u>such that</u> $||e^{-x}-\tilde{R}|| = \inf\{||e^{-x}-R||\colon R \in \tilde{R}_m\}$.

Next, we wish to turn to proving that for the m=2 case the best approximation from R_2 is actually contained in $\tilde{R}_2$. To do this we shall first show that neither of the coefficients in the denominator is zero and that the numerator and denominator do not have a common non-constant factor.

THEOREM 2.4. <u>The best approximation to</u> e^{-x} <u>from</u> R_2 <u>is not of the form</u> $\frac{ax+b}{qx+1}$, $q \geq 0$.

Proof. To prove this we use some computed results. First of all, running the Remes-Difcor algorithm as described in [2], we found the "best" approximation of the form $(ax+b)/(q_1x+q_2)$ (with $|q_1| \leq 1$, $|q_2| \leq 1$) to e^{-x} on $X = \{\frac{i}{25}\}_{i=0}^{500}$. This routine returned the values: $a = -.0934450154$, $b = .6698426328$, $q_1 = 1.0$ and $q_2 = .6330537047$. It also returned four extreme points $x_1 = 0.0$, $x_2 = .44$, $x_3 = 2.76$ and $x_4 = 20.0$ such that $e^{-x_i} - (ax_i+b)/(q_1x_i+q_2) = (-1)^i e_i$ with $e_i > .058$ for all $i=1,2,3,4$. Thus, by a de La Vallee Poussin type argument we have that $\inf\{||e^{-x}-r||_{L^\infty[0,\infty)}\colon r \in R_1^1[0,\infty)\} \geq .058$. Next, setting $r^*(x) = (a^*x+b^*)/(p^*x+1)^2$ with $a^* = -.1853243706$, $b^* = 1.022709327$ and $p^* = .524169575$ we calculated $\gamma = \max\{|e^{-x}-r^*(x)|\colon x = i/1000 \text{ for } 0 \leq i \leq 20{,}000\}$ and found that $\gamma < .023$. Next, by dividing $[0,20]$ into $[0,z]$ and $[z,20]$ where z is the zero of $r^*(x)$ we are able to show that $|E'(x)| \leq 3.2$ on $[0,20]$ where $E(x) = e^{-x}-r^*(x)$. Using this and the above value of γ with Taylor's theorem for linear polynomials we can show that $|E(x)| \leq .0246$ on $[0,20]$. Since $E(x) > 0$ and $E'(x) < 0$ for $x \geq 20$, we have that $|E(x)| \leq E(20) < .022$ for

$x \geq 20$. This completes the proof.

Note that this proof also shows that r^* is a better approximation than the one calculated in [5] for m=2. Next, we turn to proving that for any best approximation in the m=2 case, the coefficients in the denominator coalesce; that is, $q_1 = q_2$.

THEOREM 2.5. <u>Any best approximation to</u> e^{-x} <u>from</u> R_2 on $[0,\infty)$ <u>belongs to</u> $\tilde{R}_2$; <u>that is, it is of the form</u> $(ax+b)/(qx+1)^2$ <u>with</u> $q \geq 0$. <u>Furthermore</u>, $q > 0$, <u>and the numerator and denominator have no non-constant common factors</u>.

Proof. The facts that $q > 0$ and the numerator and denominator have no non-constant common factors follows from Theorem 2.4. Let $R(x) = (p_1+p_2x)/(q_1x+1)(q_2x+1)$ be a best approximation to e^{-x} on $[0,\infty)$ from R_2 with $0 < q_1 < q_2$. We first claim that $e^{-x} - R(x)$ has at least 5 alternating extreme points in $[0,\alpha]$ where α is chosen such that $x \geq \alpha$ implies $|e^{-x}-R(x)| \leq \frac{1}{2}||e^{-x}-R(x)||_{L^\infty[0,\infty)} = \frac{1}{2}\lambda_2$. This follows from the fact that $R \in R_2^1[0,\alpha]$ and has deflect zero, since if $e^{-x}-R(x)$ had fewer than 5 alternating extreme points, then the standard argument to prove alternation in $R_1^2[0,\alpha]$ R_2^1 can be used to find $\bar{R}(x)=(a+bx)/(1+cx+dx^2) \in R_1^2[0,\alpha]$ such that $||e^{-x}-\bar{R}(x)||_{L^\infty[0,\alpha]} < ||e^{-x}-R(x)||_{L^\infty[0,\alpha]}$ and with $|p_1-a|$, $|p_2-b|$, $|q_1+q_2-c|$ and $|q_1q_2-d|$ as small as desired. Thus, we can guarantee that $||e^{-x}-\bar{R}(x)||_{L^\infty[0,\infty)} < \lambda_2$ holds and that $\bar{R}(x)$ also has unequal negative roots (the discriminant of $1+cx+dx^2$ can be made arbitrarily close to that of $1+(q_1+q_2)x+q_1q_2x^2$). This, of course, is a contradiction showing that $e^{-x}-R(x)$ must have 5 alternating extreme points on $[0,\alpha]$.

Thus, $R(x)$ is the best approximation to e^{-x} on $[0,\alpha]$ from R_2^1 by the classical alternation theorem and also, therefore on $[0,\infty)$. Thus, we shall complete this proof by showing that the best approximation to e^{-x} from $R_2^1[0,\infty)$ does not have real poles. To do this, we computed the "best approximation", $R(x) = (a+bx)/(1+cx+dx^2)$ to e^{-x} from $R_2^1[0,20]$ on a 200,001 point equally spaced grid imposed on $[0,20]$. The computed results (rounded to 10 decimal places) were $a = .9911236330$, $b = -.1577830783$, $c = .6704780400$, $d = .6494291043$; the extreme points were $y_1=0$, $y_2=.2483$, $y_3=1.0852$, $y_4=3.2271$ and $y_5=13.1518$. The absolute errors at the extreme points were .0088763670 (they actually differed by less than $5\text{x}10^{-18}$) and the sign of $e^{-x}-R(x)$ was positive at y_1. The discrimant of the denominator was -2.1481756150. By direct calculation, it can be easily seen that $E(x) = e^{-x}-R(x) > 0$ and $E'(x) < 0$ for $x \geq 20$. Thus, $||E(x)||_{L^\infty[0,\infty)} = ||E(x)||_{L^\infty[0,20]}$. Now, let us assume that there exists $\bar{R} \varepsilon R_2^1$ having negative poles and for which $||e^{-x}-\bar{R}(x)|| < ||E(x)||$ holds, where for the remainder of this proof $||\cdot|| \equiv ||\cdot||_{L^\infty[0,20]}$. This will lead to a contradiction and give our desired result. We begin by noting that $|E'(x)| \leq 1$ for all $x \varepsilon [0,20]$ since $-e^{-x}$ and $-R'(x)$ have opposite signs for $x \varepsilon [0, -a/b]$ and $|R'(x)| \leq 1$ for all x since the denominator is increasing faster than the absolute value of the numerator for all x. For $x \varepsilon [-a/b,20]$ simply look at the ratio of the maximum of the numerator on this interval and the value of the denominator at $-a/b$. Thus, by the mean value theorem we have that for each $x \varepsilon [0,20]$, $|E(x)-E(\bar{x})| \leq .00005$ where $\bar{x}$ denotes a closest grid point to x. Let $\delta = .000054$, then $||E|| - \min\{|E(y_i)|: i=1,\ldots,5\} < \delta$ since $|E(y_i)-E(y_j)| < .000002$ for i, $j = 1,\ldots,5$. Since we are assuming that $||e^{-x}-\bar{R}(x)|| < ||E(x)||$, we must have that $||e^{-x}-\bar{R}(x)|| - \min\{|E(y_i)|: i \Rightarrow 1,\ldots,5\} < \delta$.

Now, there must exist i_o, $1 \leq i_o \leq 5$ such that $(-1)^{i_o}(R(y_{i_o}) - \bar{R}(y_{i_o})) > 0$ since $R \not\equiv \bar{R}$. Let us assume that $\max\{(-1)^i(R(y_i) - \bar{R}(y_i)): i = 1,\ldots,5\} = (-1)^5(R(y_5)-\bar{R}(y_5))$. Next, find $R^*(x) =$

$(a+\Delta a+(b+\Delta b)x)/(1+(c+\Delta c)x+(d+\Delta d)x^2)$ such that $R^*(y_i) = R(y_i) + (-1)^i\delta$ for $i = 1,\dots,4$. To do this we must solve the linear system $\Delta a+\Delta by_i-\Delta cy_i(R(y_i)+(-1)^i\delta)-\Delta dy_i^2(R(y_i)+(-1)^i\delta) = (1+cy_i+dy_i^2)(-1)^i\delta$, $i=1,\dots,4$. Solving this with Cramer's rule, with the determinants computed by cofactor expansion to avoid error magnification by divisions, gives $\Delta a = -.0000540000$, $\Delta b = .0004710533$, $\Delta c = -.0005063974$, $\Delta d = .0019944507$. Using this R^*, we have that $R(y_5) - R^*(y_5) = -.0000746489$ and $||R(y_i)-R^*(y_i)| - \delta| < 2 \times 10^{-18}$ for $i = 1,\dots,4$. The discrimant of R^* was -2.156832218. Now, by construction, $(-1)^i(\bar{R}(y_i) -_{>} R^*(y_i)) < 0$, $i=1,\dots,4$, and also, for $(\bar{R}(y_i) - R(y_i))<\delta$ $i = 1,\dots,5$ we must have $(-1)^i(\bar{R}(y_i)) < \delta$, since $||e^{-x}-\bar{R}(x)|| < \min\{|E(y_i)|: i=1,\dots,5\} + \delta$. Now, suppose $(-1)^5(\bar{R}(y_5)-R^*(y_5))>0$ (for, if not, then $\bar{R} \equiv R^*$ and we have our desired contradiction as R^* has non real poles). Then, we have that $\delta^* + R(y_5) = R^*(y_5) > \bar{R}(y_5)$ and that $\delta^* = (-1)^5(R(y_5)-\bar{R}(y_5)) \geq (-1)^i(R(y_i)-\bar{R}(y_i))$, $i = 1,\dots,5$ so that $|R(y_i)-\bar{R}(y_i)| \leq \delta^*$, $i = 1,\dots,5$. Letting $\bar{R}(y_i) - R(y_i) = \delta_i$, $i = 1,\dots,4$, we can estimate the coefficients of $\bar{R}$ from the equations $\bar{R}(y_i) = R(y_i) + \delta_i$, $i = 1,\dots,4$, where we know that $|\delta_i| \leq \delta^*$. Writing $\bar{R}(x) = (a+\Delta a+(b+\Delta b)x)/(1+(c+\Delta c) + (d+\Delta d)x^2)$, this system is equivalent to

$$\Delta a+\Delta by_i-\Delta cy_i(R(y_i)+\delta_i)-\Delta d_iy_i^2(R(y_i)+\delta_i)=(1+cy_i+dy_i^2)\delta_i,$$
$$i = 1,\dots,4.$$

Once again we resort to Cramer's rule to estimate Δc and Δd. Writing the determinant of the coefficients of this system as the sum of four determinants, one of which had no δ_i's in it (say D), we then computed D (= .1194079538) and estimated upper bounds for the three remaining determinants, subtracting these values from D showed that the determinant of the coefficients $\geq .1192288500$. Calculating upper bounds for the numerator determinants in the formulas for Δc and Δd and then estimating gives $|\Delta c| \leq .0012430121$,

$|\Delta d| \leq .0027602264$. Thus, letting D_1 and $\bar{D}_1$ denote the discriminant of the denominators of R and $\bar{R}$, respectively, we have that $|D_1 - \bar{D}_1| \leq .0127092755$. Treating the other cases where the maximum of $(-1)^i(R(y_i)-\bar{R}(y_i))$ occurs for i = 1,2,3, or 4, similarly, we found that $|D_1-\bar{D}_1| \leq .3959944800$, .1553098052, .0673900941 and .0219551497, respectively.

We conjecture that this result is true for all $m \geq 2$. We close this section by stating a local characterization and local uniqueness result which will be proved in a forthcoming paper[3].

Definition 2.6. $R(x) = (p_1+\ldots+p_m x^{m-1})/(qx+1)^m \in \tilde{R}_m$ *is a local best approximation to* e^{-x} *on* $[0,\infty)$ *if there exists a* $\delta > 0$ *such that if* $\bar{R}(x) = (\bar{p}_1+\ldots+\bar{p}_m x^{m-1})/(\bar{q}x+1)^m \in \tilde{R}_m$, $|p_i-\bar{p}_i| < \delta$, $i=1,\ldots,m$ *and* $|q-\bar{q}| < \delta$ *then* $||e^{-x}-R(x)|| \leq ||e^{-x}-\bar{R}(x)||$. *If, in addition, strict inequality holds whenever* $\bar{R}(x) \not\equiv R(x)$ *then* R *is said to be locally unique*.

THEOREM 2.7. *Let* $m > 1$. *Then a nondegenerate* $R(x) = P(x)/Q(x) = (p_1+\ldots+p_m x^{m-1})/(qx+1)^m \in \tilde{R}_m$ (i.e. $R \not\equiv 0$, $P(x)$ *and* $Q(x)$ *have no common factors and* $q > 0$) *is a best local approximation to* e^{-x} *from* $\tilde{R}_m$ *on* $[0,\infty)$ *if and only if* $e^{-x} - R(x)$ *has at least* $m + 2$ *alternating extreme points*. *Whenever this occurs*, R *is locally unique*.

We remark that numerical examples seem to suggest that there exist distinct R_1, $R_2 \in \tilde{R}_2$ satisfying this theorem.

3 Numerical Results

Our initial algorithm for computing approximations to e^{-x} from R_m and $\tilde{R}_m$ involved linearizing the denominator by Taylor's Theorem and setting up an iterative procedure, using the differential correction algorithm to compute an approximation at each inner stage. Precisely, for R_m set $g(q_1,\ldots,q_m,x) = \prod_{i=1}^{m}(q_i x+1)$

and define $\psi_j(q_1,\ldots,q_m,x) = x\prod_{\substack{i=1\\ i\neq j}}^{m}(q_ix+1)$ for $j=1,\ldots,m$, $\psi_o(q_1,\ldots,q_m,x) = g(q_1,\ldots,q_m,x) - \sum_{\nu=1}^{m} q_\nu\psi_\nu(q_1,\ldots,q_m,x)$. Thus, if $\bar{R}(x) = \bar{P}(x)/\prod_{i=1}^{m}(\bar{q}_ix+1)$, $0 \leq \bar{q}_1 \leq \bar{q}_2 \leq \ldots \leq \bar{q}_m$ is an approximation to e^{-x} at some step in the algorithm, then a new approximation $R(x) = (p_o+p_1x+\ldots+p_{m-1}x^{m-1})/\prod_{i=1}^{m}(q_ix+1)$ is found by calculating $p_o,\ldots,p_{m-1},q_1,\ldots,q_m$ that minimize $||e^{-x}-(p_o+\ldots+p_{m-1}x^{m-1})/(q_1\psi_1(\bar{q}_1,\ldots,\bar{q}_m,x)+\ldots+q_m\psi_m(\bar{q}_1,\ldots,\bar{q}_m,x)+\psi_o(\bar{q}_1,\ldots,\bar{q}_m,x))||$ over T a finite subset of $[0,N]$. Observe that the denominator in this problem is precisely the linearization of $g(q_1,\ldots,q_m,x)$ via Taylor's Theorem applied to the first m independent variables. This minimum can be calculated by the differential correction algorithm. Since this is a linearization of the problem we wish
→ m to solve, if we force an ordering on the $q_1,\ldots,q_n$ to get a dego to n
unique solution, it seems reasonable to expect that if the initial approximation is sufficiently close to a best approximation then this algorithm will converge to that best approximation.

This approximation must be calculated on a large interval (the length of the interval needed seems to increase as a function of m, but not monotonically) to give a candidate for a best approximation to e^{-x} on $[0,\infty)$; and since we wish to get an accurate approximation of the continuous solution, we must use a fairly fine mesh so that card (T) will be large. Since the differential correction algorithm tends to become unstable as card (T) grows large, we decided to use the Remes-Difcor algorithm [2] for calculating the linearized minimum. We did this because this algorithm applies the differential correction algorithm to certain (small) subsets of T chosen in such a manner (depending upon alternation) that convergence to the solution on the whole space occurs. Thus, we had no a priori guarantee that this would work since a standard alternation theory does not exist for the linearized minimization problem due to the addition of the constraints on $\{q_i\}_{i=1}^{m}$. However, in spite of this, the results of the algorithm are acceptable in that the algorithm returned (or tried to return) a solution in which the q_i's coalesced and for which the error curve $e^{-x} - P(x)/\prod_{i=1}^{m}(qx+1)^{m}$ (the final coalesced approximation) alternated on m+2 points of T. Thus, by an alternation theorem we have proved [3], we have a best local approximation from $\tilde{R}_m$ and as we conjectured earlier; therefore, also from R_m; we also ran an algorithm of this character for the class $\tilde{R}_m$. In all cases it has given the same results as the above algorithm applied to R_m. A precise study of these algorithms remains to be done and we conjecture that convergence results can be proved for both R_m and $\tilde{R}_m$, at least using the differential correction algorithm for the inner minimization.

We have run these algorithms for various values of m using a grid with spacing .002 imposed on an interval $[0,N]$, where N is chosen by trial and error so that the computed results make it apparent that the error norm on $[N,\infty)$ is smaller than the error on $[0,N]$. The computations were done on a UNIVAC 1106, which has roughly 18 digits of accuracy in double precision. Initially, we start with $\bar{p}_o=1$, $\bar{p}_1=\ldots=p_{m-1}=0$, $\bar{q}_j = \frac{j}{m}$, j=1,...,m and ran the program with additional constraints $q_i \leq q_{i+1}$ - DIFF where DIFF is a nonnegative parameter. If DIFF > 0 we found that the computed q_i's immediately differed by exactly DIFF, and if DIFF was set equal to 0 the algorithm ran and the computed q_i's coalesced. The algorithm for $\tilde{R}_m$ had a linearization in which the denominator of the approximation is $g \cdot m\, x(\bar{q}x+1)^{m-1} + [(1-m)\bar{q}x+1](\bar{q}x+1)^{m-1} \equiv q\psi_1(\bar{q},x) + \psi_o(\bar{q},x)$ where $\bar{q}$ is the value from the previous approximation. Here the initialization was $\bar{p}_o = 1$, $\bar{p}_1 = \ldots = \bar{p}_{m-1} = 0$, $\bar{q} = 1/n$. Although we allowed this program to run for seven outer iterations, the coefficients nearly always stopped changing after four or five outer iterations,

→ ch. g → to q.

and the computed absolute values of the errors at the m + 2 extreme points agreed to at least fourteen significant figures. The results are shown in the table below, with the error of [5] given in the last column for comparison purposes. The sign attached to the last extreme point is the sign of E(x) at that point. It should be noted that it is possible (although unlikely) that in some cases there is a local best approximation other than ours which gives a smaller error. In the m = 3 case we have found another local best approximation (with g = 1.05109 and ||error|| = 1.33720 (-02) and in the m = 6 case there appear to be at least three local best approximations other than the one in the table.

Finally, we would like to thank Professor R.S. Varga for bringing [4] to our attention where some of the results of this paper and of [3] have also been obtained independently.

Table of Numerical Results

m	last ext. pt.	q	||error||	||error|| [5]
2	12.932+	.52416	2.27093 (-02)	2.49038 (-02)
3	37.250-	.27127	8.04713 (-03)	1.5053 (-02)
4	83.814+	.17797	3.30771 (-03)	7.85325 (-03)
5	80.802+	.27866	1.16064 (-03)	3.05486 (-03)
6	152.352-	.19296	4.26252 (-04)	8.89316 (-04)

References

1. Cody, W.J., G. Meinardus and R.S. Varga, Chebyshev rational approximations to e^{-x} on $[0,\infty)$ and applications to heat-conduction problems, J. Approximation Theory, 2 (1969), 50-65.

2. Kaufman, E.H., Jr., D.J. Leeming and G.D. Taylor, A combined Remes-Differential correction algorithm for rational approximation, submitted.

3. Kaufman, E.H., Jr. and G.D. Taylor, Uniform approximation with rational functions having negative poles, submitted.

4. Lau, T. C-Y., Rational exponential approximation with real poles, preprint.

5. Saff, E.B., A. Schönhage and R.S. Varga, Geometric convergence to e^{-z} by rational functions with real poles, Numer. Math., Vol. 25 (1976), 307-322.

6. Saff, E.B. and R.S. Varga, Angular overconvergence for rational functions converging geometrically on $[0,+\infty)$, Theory of Approximation with Applications (edited by Law and Sahney) Academic Press, New York, 1976, 238-256.

E.H. Kaufman, Jr.
Department of Mathematics
Central Michigan University
Mount Pleasant, Michigan 48859

G.D. Taylor*
Department of Mathematics
Colorado State University
Fort Collins, Colorado 80523

*Research sponsored by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under Grant No. AFOSR-76-2878.